(54) Title: MEASUREMENT OF SPEECH SIGNAL QUALITY
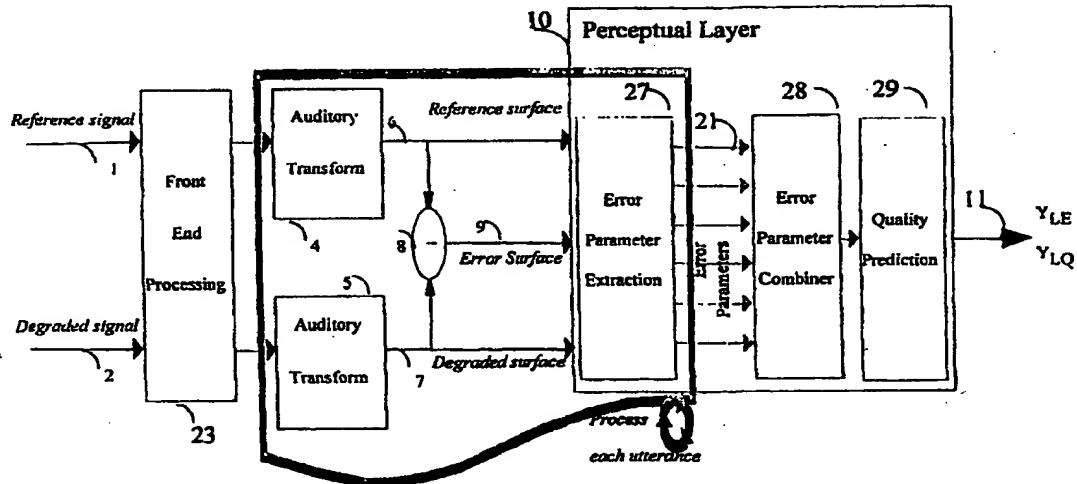
(57) Abstract

Apparatus for assessing the performance of telecommunications systems by comparison of a reference signal with the same signal as degraded by the system under test tends to give unreliable results when used to test connectionless packet–switching equipment. This is because variable delay in delivery of the individual packets can result in the test signal and reference signal becoming unsynchronised. This may lead to the apparatus recording a poorer performance than a human tester would have recorded. To overcome this problem, the invention separates the test signal into a series of individual sections, which are synchronised and analysed separately, and a quality value, aggregated over all the sections, is derived.

MEASUREMENT OF SPEECH SIGNAL QUALITY

This invention relates to the measurement of quality of a sound signal, and more specifically a speech signal. Objective processes for this purpose are currently under development and are of application in prototype testing, pre-delivery testing of components, and in-service testing of installed equipment. They are most commonly used in telephony, but are also of application in other systems used for carrying speech signals, for example public-address systems.

The present applicant has a number of patents and applications relating to this technical field, most particularly European Patent 0647375, granted on 14th October 1998. In this system, a signal degraded by the system under test is compared with a reference signal, which has not passed through the system under test, to identify audible errors in the degraded signal. These audible errors are assessed to determine their perceptual significance – that is, errors of types which are considered significant by human listeners are given greater weight than are those which are not considered so significant. Since only audible errors are assessed, inaudible errors, which are perceptually irrelevant, are not assessed.

The automated system provides an output comparable to subjective quality measures originally devised for use by human subjects. More specifically, it generates two values, Y_{LE} and Y_{LQ}, equivalent to the "Mean Opinion Scores" (MOS) for "listening effort" and "listening quality", which would be given by a panel of human listeners when listening to the same signal, as will be discussed later. The use of an automated system allows for more consistent assessment than human assessors could achieve, and also allows the use of compressed and simplified test sequences, and multilingual test sequences, which give spurious results when used with human assessors because such sequences do not convey intelligible content.

Such automated systems require a known (reference) signal to be played through a distorting system (the telephone network) to derive a degraded signal, which is compared with an undistorted version of the reference signal. Such systems are known as "intrusive" measurement systems, because whilst the test is carried out the system under test cannot carry live (revenue-earning) traffic.

An auditory transform of each signal is taken, to emulate the response of the human auditory system (ear and brain) to sound. The degraded signal is then compared with the reference signal in the perceptual domain, in which the

subjective quality that would be perceived by a listener using the network is determined from parameters extracted from the transforms.

A suitable test signal is disclosed in International Patent Specification WO/95/01011 (EP0705501) and comprises a sequence of speech-like sounds,
5   selected to be representative of the different types of phonetic sounds that the system under test may have to handle, presented in a sequence. The sounds are selected such that typical transitions between individual phonetic elements are represented. Typical speech comprises a sequence of utterances separated by silent periods, as the speaker pauses to breathe, or listens to the other party to the
10   conversation. These silent periods, and the transitions between utterances and silent periods, are also modelled by the test signal.

This existing system reliably assesses most speech carrier technologies employed within conventional analogue and digital switched telephone networks. In such networks a dedicated connection is provided between the two parties to a
15   call, for the duration of that call, and all speech is carried over that connection. However, connectionless packet-based speech transmission systems are beginning to be introduced, in particular for use in the "Internet" and companies' internal "Intranets". In a connectionless packet-based system each transmission is divided into a series of data "packets", which travel independently from one user to the
20   other. Intermediate nodes in the network transmit the packets to each other according to address information carried in each packet. However, according to the demands of other traffic on the various links between such nodes, and the available capacity on those links, different packets may be delayed, or may travel by different routes to reach the same destination. Consequently, end-to-end times
25   vary from one packet to another. For the transmission of data such as text, or downloading of computer files for the recipient to use subsequently, such variations in end-to-end times are of little consequence. However, when used for real-time speech, these variations can affect the clarity of the speech as perceived by the user.
30   Various proposals have been made to try to minimise the delay to a level which does not interfere with conversation and comprehension – see for example the present applicant's International Patent Application WO99/12329, and the article by R Barnett in *"Electronics and Communication Engineering Journal"*, October 1997, entitled *"Connectionless ATM"*. However, it is fundamental to such

connectionless systems that some variation in the residual delay will occur. A single speech utterance is typically assembled from the information carried in several packets. However, variations in the delay between individual packets will in general not be apparent in the resulting utterance, as the slowest packet generally

5   determines the delay to the utterance as a whole. However, the delay to each complete utterance can vary considerably between one utterance and the next, as buffer lengths are normally adjusted during periods of silence.

Changes to the delay occurring during the course of an utterance, for example because part of the utterance is missing, will be more apparent in the

10 resulting utterance.

In addition to changes in residual delay, transmission systems are now beginning to come into use in which changes in other characteristics, such as level (signal amplitude), can occur. See ITU-T draft recommendation G.169.

The human brain is insensitive to small changes in delay and amplitude

15 between speech events, so these variations may be imperceptible to a human listener, provided the magnitude of the effect is not such as to interfere with conversation. However, the prior art measuring system is sensitive to such variations, so that it returns unreliable values for signal quality when testing connectionless packet systems – that is, the results do not accurately reflect the

20 subjective quality reported by human subjects.

If the delay is constant, the two signals can easily be synchronised to take account of the delay. However, if the degraded signal suffers variable delay, at least some parts of the degraded signal would not be synchronised with the test signal. The lack of synchronisation in those parts would be detected as substantial

25 errors, which would be so great as to mask any errors caused by actual degradation of the signal. This would lead to an inaccurate measure of the subjective effect of the degradation.

There is therefore a requirement for a measurement system that is robust against such variable delays.

30 According to the invention, there is provided apparatus for testing equipment for handling speech signals, comprising

means for receiving first and second signals, means for selecting individual sections in the first signal and second signal,

means for comparing each section in the second signal with the corresponding section in the first signal to generate a distortion perception measure which indicates the extent to which the distortion of said section would be perceptible to a human listener, and

5   means for combining the results of each such measurement to generate an overall measure of the extent to which the distortion of the second signal with respect to the first signal would be perceptible to a human listener.

Preferably, the overall measure takes account of the perceptual importance of each section. The perceptual importance of a given section will depend on the
10   number of individual speech components, and their relative importance to subjective quality measures, in that section.

The means for selecting individual sections in the two signals may comprise means for identifying individual utterances. In the preferred embodiment this is achieved by detecting the end of each silent period in the signal. The apparatus
15   preferably includes means for synchronising each section in the distorted signal with the corresponding section in the test signal. Synchronisation is preferably carried out by analysis of the speech content of the signals. However, a separate synchronisation characteristic may be used to identify the onset of each section. This synchronisation characteristic is preferably outside the frequency band
20   characteristic of speech, so that it does not interfere with the analysis process (which only detects changes perceptible to a human listener). The synchronisation characteristic related to a given section may be chosen to be unique to that section, to ensure that each distorted section is compared with the corresponding test section. This ensures that, should a section, or its synchronisation
25   characteristic, be lost as a result of the distortion, subsequent sections can nevertheless be analysed.

In a preferred arrangement, each section is analysed to identify the position of any delay change, and the parts of the section preceding and following any such delay change are separately synchronised, and analysed for degradation.

30   In the embodiment to be described in detail, the sections which are selected for analysis may comprise individual utterances, that is   unbroken sections of speech each preceded and followed by silence of a minimum pre-determined length.   However, a number of alternative methods may be used for defining suitable sections. For example, long utterances as previously defined may be sub-

divided into two or more sub-utterances. The signal may instead be broken into a number of sections of fixed length or a fixed number of equal-length sections. However, if any sections contain no speech at all, they are preferably not used for analysis as delay is harder to determine. Any errors in non-information containing

5   sections are less likely to be perceptually important.

In another aspect, the invention comprises a method of testing equipment for handling speech signals, comprising the steps of

supplying a test signal,

receiving a distorted signal which corresponds to said test signal when

10   distorted by equipment under test,

selecting individual sections in the test signal and distorted signal, comparing each section in the distorted signal with the corresponding section in the test signal to generate a distortion perception measure which indicates the extent to which the distortion of said section would be perceptible to a human

15   listener, and

combining the results of each such comparison to generate an overall measure of the extent to which the distortion of the signal would be perceptible to a human listener.

The invention may be embodied in computer software, as a computer

20   program product for loading directly into the internal memory of a digital computer, comprising software code portions for performing the steps of the method described above when said product is run on a computer.

In a further aspect the invention comprises a computer program product stored on a computer usable medium, comprising:

25   computer-readable program means for causing a computer to select individual sections in a first signal and in a second signal,

computer-readable program means for causing the computer to analyse each section in the first signal and the second signal to generate a distortion perception measure which indicates the extent to which the distortion of said

30   section in the second signal as compared with the first signal would be perceptible to a human listener, and

computer-readable program means for causing the computer to combine the results of each such comparison to generate an overall measure of the extent to which the distortion of the signal would be perceptible to a human listener.

The computer program product may be embodied on any suitable carrier readable by a suitable computer input device, such as CD-ROM, optically readable marks, magnetic media, punched card, or on an electromagnetic or optical signal.

A preferred embodiment of the invention will now be described, by way of
5 example, with reference to the drawings, in which:

Figure 1 shows, in diagrammatic form, the general arrangement of the prior art system;

Figure 2 shows, in diagrammatic form, the general arrangement of a system according to the invention;

10 Figure 3 shows, in more detail, the arrangement of the pre-processing elements of Figures 1 and 2;

Figure 4 shows, in more detail, the arrangement of the auditory transform elements of Figures 1 and 2;

Figure 5 illustrates the utterance identification process used in the
15 preferred embodiment of the invention;

Figure 6 illustrates the utterance alignment process used in the preferred embodiment of the invention; and

Figure 7 illustrates the functional elements of a general-purpose computer suitable for performing the invention.

20 Figure 8 illustrates a modified process according to the invention, allowing a single utterance to be aligned in two parts

Figure 9 is a time line illustrating the process shown in Figure 8

The functional components illustrated in these figures can be embodied as software running on a general-purpose computer.

25 The prior art system shown in Figure 1 incorporates two inputs 1,2 for providing the reference signal and the degraded signal. These undergo pre-processing in a front end unit 3, primarily in order to synchronise the two signals, as will be described in more detail with reference to Figure 3. They then undergo auditory transformation in a transform unit 4,5 to model the non-linear response of
30 the human auditory system to frequency, and the frequency-dependent response of the auditory system to amplitude. These transforms will be described in more detail with reference to Figure 4. They each generate a respective output 6,7, comprising a series of values representing the signal as a set of frequency bands

and of time intervals, which can be represented as a two dimensional matrix, or graphically as a surface.

The frequency bands are of unequal width, to allow for the non-linear response of the human auditory system. The time intervals need not be the same

5   in each auditory band, as discussed in European patent EP 0647375.

These two outputs 6,7 are compared by a comparator 8 which generates an output 9 which is a measure of the difference between them. This output 9 represents the audible error applied to the reference signal 1 by the system under test in order to generate the degraded signal 2. The three outputs 6,7,9 are then

10 processed by a perceptual analysis unit 10 which determines the perceptual significance of each audible error identified on the output 9. This process takes into account contextual factors such as whether the error is masked by other frequency or time components, whether the error is of a type which affects comprehension, and other factors discussed in detail in the specifications referred

15 to previously. Signals are then generated at the output 11, which give an overall value for listening effort (Y_{LE}) – a measure of how much a listener would need to concentrate to extract meaning from a signal carrying such distortions, and for listening quality (Y_{LQ}) – the subjective quality of the signal. Both scales rate the signal on a scale from 1 to 5 (larger numbers signifying less degradation), and are

20 described in ITU-T P.800.

Figure 3 shows schematically the processes that take place in the front-end unit 3. Each signal is first subjected to input conditioning (units 12/13) to remove simple systematic differences such as DC level. The signals are then passed to filters 14/15, which emulate the frequency response between the point

25 of signal acquisition and the inner ear. Each signal is then passed to a respective speech detection stage 16/17, to identify the onset of speech, or some other predetermined cue in the signal. This allows corresponding parts of the reference signal and degraded signal to be identified in an alignment processor 18 so that the two signals can be processed to generate an error surface. In the prior art system

30 of Figure 1, which is configured to assess the quality of switched connections in which delay is constant, this alignment is sufficient to synchronise the entire speech sample. However, in the present invention individual packets, and individual utterances assembled from those packets, are subject to variable delays around the value determined by the alignment processor 18.   In   general,   the

degraded signal will carry some degree of attenuation or amplification as compared with the original reference signal. Simple attenuation or amplification, if within the limits of audibility, will not affect the perceptual quality of the speech content, so such changes in level must be allowed for in the processing. The alignment

5   processor 18 compares the absolute level of the two cue signals, so that attenuation can be compensated for in the comparison.

Figure 4 shows the functional elements of the auditory transform units 4,5. The signal is first filtered 19 into a number of separate frequency bands, each of which is then averaged (20) in the time domain over an interval. The interval

10 may be different for different frequency bands, according to considerations such as the duration of masking effects, and the audibility of short-duration sounds, which effects vary with frequency. The psycho-acoustic loudness of each time and frequency component, taking into account masking effects, frequency, and other psychological and physiological factors, is then determined (21). Values below an

15 audibility threshold value are then eliminated (22) to generate the outputs 6,7.

The auditory transform units 4,5, and the components 12,13,14,15,16,17 of the front end processor 3 are shown as being duplicated, to allow the reference and degraded signals to be processed in parallel. It is possible to carry out these processes sequentially for the two signals, using a single input conditioning unit,

20 filter and speech detector to process first the reference signal, and then carrying out the same process on the degraded signal. Both outputs are then applied to the alignment processor 18. Similarly, the two reference surfaces 6,7 may be generated by the same auditory transform unit. Clearly, the results of the first processing have to be stored whilst the second signal is processed.

25 In the prior art model speech and non-speech are identified in the speech detectors 16, 17, to select and align the complete samples for use in the perceptual analysis unit 10. In contrast, in the embodiment of the invention shown in Figure 2, the perceptual analysis unit 10 operates on a single section at a time. In the invention, alignment, level and transfer function are estimated independently

30 for each utterance. Multiple sections within a test signal are processed individually and their extracted parameters then combined.

The inputs 1,2 are applied to a front end processor 23, which carries out similar functions to the processor 3 in Figure 1, but operates to select and align each individual section instead of simply aligning the signal as a whole. This will be

described further, with reference to Figures 5 and 6. The processor 23 generates a series of aligned section pairs.

The auditory transform units 4,5 and error parameter extraction unit 27 then handle each section pair, in the same way as the prior art process, to 5   generate error parameters for each utterance, at the output 21.

These error parameters are then further processed in a combiner unit 28, which generates, for each error parameter, an overall value aggregated over all the sections. These overall values are typically an aggregate or average value of the individual section parameters, weighted to take account of the relative perceptual 10   significance of each section, for example its length. These weightings can be determined in advance, since the test signal is a predetermined reference signal. The error parameter combiner 28 ensures that errors occurring in different sections are appropriately aggregated. For example the length of a particular section would influence its impact on the quality. Also the combiner may include an error 15   aggregation profile — errors occurring at the beginning or end of a call may have greater impact on perceived quality than those in the middle of the call.

Finally, the combined error parameter is converted to the standard ITU signal quality scale in a translation unit 29, to generate values of Y_{LE} and Y_{LQ} at the output 11.

20   A process carried out in the front end processor 23 to select the sections for analysis will now be described in detail with reference to Figure 5, which shows parts of a reference signal 30, and a degraded version 40 of that signal. In this process individual utterances are identified. An utterance is regarded as a period of speech in which an instance of non-speech does not exceed a defined 25   threshold duration. As shown in Figure 5, the reference signal has two periods of silence, 31, 32. The first period, 31, is longer than the threshold value and thus defines a gap between two utterances 33, 34. The second period, 32, is shorter than the threshold and thus does not define a gap between two utterances. Some voice-over-packet switching systems are arranged to ensure that each packet of a 30   transmission must be routed the same way as the previous packet unless a predetermined number of preceding packets have all been silent. This ensures that all packets of any one utterance are routed the same way, although different utterances can be routed differently, the silent gaps being consequently shorter or longer at the receiving end. The threshold used in the present embodiment may

conveniently be selected to correspond to this predetermined number of silent packets.

To maintain robustness the beginning and end marks for the utterances 43, 44 in the degraded signal are taken only from the reference signal 30. This
5   ensures that audible interference on the degraded signal 40, or loss of the signal, does not interfere with the identification of the utterances. For example, in Figure 5 the last part 45 of the second degraded utterance 44 has been lost, however the end 46 of the degraded utterance 40 is located at the point corresponding to the end 36 of the reference utterance 30.

10 The human brain is insensitive to small changes in delay or level between speech events. Processing utterance by utterance allows these small changes to be disregarded. Figure 6 illustrates this process.

It is assumed that the front end processing module 3, (Figure 3), has accounted for the bulk delay offset between the reference signal 50 and degraded
15 signal 60. With the bulk offset removed each utterance 61, 62 in the degraded signal 60 is expected to be within an align-window of n samples either side of the corresponding reference utterance 51, 52. A portion of the reference utterance 51, herein referred to as an utterance search area 53, is processed by correlating it with the corresponding portion 63 of the degraded signal, using a cross-correlation
20 method on overlapping parts. The method may be that used in the applicant's European Patent 0791251. The mode of the correlation function identifies the precise start point of the degraded utterance 61. As a result the degraded utterance 61 can be aligned with the reference utterance 51, allowing this portion to be processed through the rest of the analysis unit. The process is repeated for
25 subsequent utterances 52, 62, etc.

Level alignment, transfer function estimation, auditory transforms and error parameter extraction can then all be performed on an utterance by utterance approach.Figure 7 illustrates schematically a general-purpose computer suitable for performing the invention. It comprises an input device 71, a central processing unit
30 72, an output device 73 such as a printer or visual display unit, a memory 74, and a port 75 for connection to other devices. The operating instructions for controlling the computer may be supplied in machine-readable form on a carrier such as a magnetic disc or tape 70, or a signal downloaded through the port 75 from another device by way of a telecommunications connection. Similarly, the reference signal

and degraded signal may be supplied by either means. The operating instructions and the reference and degraded signals are stored in the memory 74 until called up by the central processor 72, which operates in accordance with the instructions carried in the program, and any human inputs made through an input device
5   connected to the port 75. The output parameters, Y_{LE} and Y_{LO} are supplied to the output device 73.

It will be noted that differences in delay within an individual utterance (speech event) may also occur. Differences in delay between one part of an utterance and another part of the same utterance are more noticeable than
10   differences in delay between separate utterances, so they are more likely to affect the subjective quality of the signal and it is more important to account for them. In particular, a reduction of the delay between the original and degraded signal during an utterance will be apparent as the shortening of an individual sound, or its complete absence. Similarly, an increase in the delay between the original and
15   degraded signal during an utterance will be apparent as the lengthening of an individual sound, or the introduction of a sound absent from the original signal. These errors have an effect on the perceived quality of the signal. However, the misalignment of the signals before or after the change in delay would also be perceived by the measuring system as degradation, even though no degradation
20   would be detectable to the listener.

In a modification of the invention, which will now be described with reference to Figures 8 and 9, such discontinuities in delay may be identified and allowed for. This process will identify an utterance with a single such discontinuity, unless it is close to the beginning or end of the utterance. It will also
25   identify some utterances with two such discontinuities.

This process estimates for a given pair (original/degraded) of speech files any temporal delays present in the degraded file and the locations of any delay changes. The process is illustrated in Figures 8 and 9.

In an initial processing step the original signal 1 and degraded signal 2 are
30   read (step 801). If the degraded signal is shorter than the original signal, zero padding is added to the end of the degraded signal (step 802). If the degraded signal is longer, its length is adjusted by truncating it. This ensures the integrity of the original signal is not affected.

Next, areas of speech activity (utterances) are located (step 803) using a voice activity detector; for each utterance from the original file the section with the same sample labels from the degraded file is identified (step 804). For each such utterance pair the following steps are then performed.

5   Since the voice activity detector produces cut-off points for the extracted speech sections directly where speech starts and ends, there is no run-in or run-out; a short period of silence is therefore added to each end of the two utterances in order to aid the alignment process (step 805).

The next step is to calculate a delay estimate for the complete utterance

10   (step 806) and store the result. A confidence value for this delay estimate is also calculated and stored.

The utterance is next split into two sections at the midpoint m (see Figure 9), and delays and confidence values for each half are calculated and stored (step 807).

15   If the delays in the two halves match (within a small error margin), this is indicative that the delay is constant throughout the utterance and there is no delay change. The following trisection procedure 808-826 is then used to produce a further confidence measure to confirm this result, using the delay value already determined (step 806). If the two delays do not match (within a small error

20   margin), this is indicative that the second section has an average delay different from that in the first, indicating that a delay change is present in one of the two halves. The following trisection procedure 808-826 is used to determine the position of the delay change.

In this specification, two delay estimates are said to "match" if they differ

25   by less than a predetermined tolerance, for example 16 digitisation samples (1 millisecond).

Note that delay changes very near the beginning or end of the utterance will not be detected, because the two halves will then have delay values that are very similar. In such circumstances the original and degraded utterances will be

30   aligned correctly except for a short section at one end, which will be identified as a degradation.

The trisection procedure is an iterative process carried out on a search interval [a,b] which is initially the complete utterance (as shown in Figure 9) but is re-defined on each iteration. The search interval is analysed in three symmetrical

overlapping sections of equal length, as shown in Figure 9, each of the three sections comprising 40% of the search interval. They are defined by the intervals: [a,x], [w,z], [y,b], where a, w, x, y, z, b are respectively at 0.0, 0.3, 0.4, 0.6, 0.7, and 1.0 fractional distances of the complete search interval [a,b]. As they form a

5   symmetric set about the midpoint m (0.5 fractional distance from either end of the search interval), the search direction is unbiased.

To begin the iterative trisection process, the search interval [a,b] is set to be the entire sample (step 808).

At each step of the iterative process, the three overlapping intervals [a,x],

10   [w,z] and [y,b] are defined as the first, middle, and last 40% of the search interval respectively (step 809). Delays and corresponding confidences are then measured and stored for each section (step 810).

It has already been noted that if a delay change occurs close to the beginning or end of a search interval, it will have a relatively small effect on the

15   delay value estimate for that interval, and is unlikely to be detected as a delay change. Thus delay changes near the beginning or end of an utterance are less detectable than those nearer the middle. A delay change occurring for example in the region [w,x] (Figure 9) will result in the delay value estimated for the first interval [a,x] being influenced mainly by the actual delay before the change, but

20   the delay value estimated for the middle interval [w,z] will closely correspond with the actual delay after the change. This second delay value will also be measured, but with a greater confidence value, for the third interval [y,b]. Thus the delays for the second and third intervals [w,z] and [y,b] will match, in the sense defined above.

25   In the next stage, the delays measured for the three intervals are compared (step 811)

If the first and third intervals [a,x] and [y,b] match, the search is terminated and the iterative loop exited (step 812), as it is to be concluded in this case that there is no delay between the beginning and end of the search interval.

30   The delay for the complete search interval, and a confidence value, are then estimated and stored, and the iterative loop exited (see step 816).

If none of the three intervals match (step 813), the search is terminated and the iterative loop exited (step 816) since in this case it is not possible to

determine a direction for the next iteration. This situation will occur if two or more delay changes are present.

If the delay for the middle interval [w,z] matches one (but not both) of the other two [a,x], [y,b] (step 814), the delay change can be assumed to lie within or near the overlap between the second interval [w,z], and the non-matching interval [y,b] or [a,x]. The search interval is then redefined to become [a z] or [x,b] (step 815) (eliminating the interval within which the delay change does not occur) and another iteration is made (step 809).

The iterative procedure terminates when the length of the search interval decreases below a predetermined threshold, (for example 8000 samples: 0.5sec), (step 816) unless the loop has already been exited (steps 812, 813).

The delay values and confidence values already stored (steps 806, 807, 810, 812) are now used to determine the changeover point and actual delay values.

Firstly, the stored confidence value for each iteration of the procedure, including the full utterance and half-section stages, are retrieved, and the mean confidence value for each iteration is determined (step 817). The set of delay values corresponding to the maximal mean confidence is then selected (step 818) as this will be the set in which the estimated position of the delay change (if any) is closest to the exact value.

To identify the bounds within which the actual changeover point falls, the following steps are performed on the selected set of delay values:

If the delay calculated for the whole utterance alignment (step 806) gives the delay with the largest confidence, it is inferred that there is no delay change and so no changeover point (step 819).

If the largest mean confidence value is identified as that obtained for the two delays calculated for division of the utterance into equal halves (step 807), and those halves match, then again there is no delay change and so no changeover point (step 820). If the halves do not match, then there is a delay change, the changeover point being identified as exactly the midpoint of the utterance (step 821).

If the set of data values selected as having the highest confidence value is that of one of the trisections, the following steps 822 – 826 are followed.

If the best confidence value is for an iteration in which the first and third sections match (as found in step 812) then it is assumed that there is an error in calculating the middle section delay, (which is harder to obtain an accurate delay estimate for, since there are no beginnings and ends of speech to align) and that 5 no delay change has actually occurred. (Step 822) The alternative possibility, that two complementary delay changes have occurred at exactly the right time such that the outer intervals [a,x], [y,b] match, with a different delay in the middle section [w,z], is too unlikely to require consideration.

If the delay in the middle interval [w,z] of the selected trisection matches 10 one (but not both) of the other two intervals (as found in step 814), there must be a delay change within the search interval [a,b] (step 823). In particular, if the delay estimates match for the first and second intervals [a,x], [w,z], then the delay change must lie within the second half of the second interval [w,z], or the first half of the third interval [y,b]. Similarly, if the delay estimates match for the second 15 and third intervals [w,z], [y,b] then the delay change must lie within the second half of the first interval [a,x], or the first half of the second interval [w,z]. The exact position of the changeover cannot be determined more precisely than the minimum search interval allows, so a predetermined point falling within the identified range of the selected search interval is selected as the estimated delay 20 change position (step 825). (Note that the selected search interval itself is typically a small fraction of the complete utterance, as a result of the iterative reductions in the search interval (step 815), so the location of the delay change will be identified as falling within that short interval).

If no sections match (step 824), it is assumed that there are two delay 25 changes (more than two are highly unlikely, and two predetermined points along the search interval are selected as the estimated delay change positions (step 825).

New values for the delay in each part of the utterance, as now defined by the estimated delay change point or points, are then determined (step 826).

30 Depending on the outcome of the selection step 818, the utterance may therefore have no, one or two delay changes identified. The original utterance pair 1,2 is then separated into separate sections at these delay changes (if any) and the delays so identified imposed on each pair (step 827).

The process (steps 805-827) is repeated for the next utterance pair.

In the embodiment described so far, the section initially selected is an individual utterance. Utterances of duration longer than some pre-determined threshold may be sub-divided into two or more sub-utterances, each sub-utterance then being aligned separately, thereby allowing a plurality of delay changes to be

5   identified in a single utterance.

The presence of a delay change may be tested by dividing the utterance in two at one or more different points. For each division point, the two parts of the signal are aligned separately and the confidence of such alignments is determined. A comparison across all division points of the delays of each part and the

10   confidence of these delay values is used to determine the location, if any, of a delay change in the signal. This process may then be repeated for the sections on either side of this identified delay change to identify possible additional variations in delay.

In another alternative arrangement, the complete signal (or each of several

15   predetermined sections of it such as individual utterances) is first processed as a whole to identify and cancel any constant delay. The signal is then divided in two and alignment performed on each section separately. Each section so aligned is recursively sub-divided until some pre-determined minimum duration is reached. This process enables a number of delay changes at arbitrary locations in the signal

20   to be identified.

CLAIMS

1.   Apparatus for testing equipment for handling speech signals, comprising means for receiving first and second signals,

5   means for selecting individual sections in the first signal and second signal,

means for comparing each section in the second signal with the corresponding section in the first signal to generate a distortion perception measure which indicates the extent to which the distortion of said section would be perceptible to a human listener, and

10   means for combining the results of each such measurement, to generate an overall measure of the extent to which the distortion of the second signal with respect to the first signal would be perceptible to a human listener.

2.   Apparatus according to claim 1, wherein the combining means derives an

15 overall measure, which takes account of the perceptual importance of each section.

3.   Apparatus according to claim 1 or claim 2, wherein the means for selecting individual sections in the signals comprises means for detecting the end of each

20 silent period exceeding a predetermined length, thereby detecting individual utterances in the signal.

4.   Apparatus according to claim 3, wherein an utterance containing more than a pre-determined length of speech is sub-divided into two or more sections.

25

5.   Apparatus according to claim 1 or claim 2, wherein the means for selecting individual sections in the signals comprises means for detecting whether speech is present in each of a number of arbitrarily chosen sub-divisions of the signal.

30 6.   Apparatus according to any preceding claim, wherein the means for analysing each section includes means for synchronising the section in the second signal with the corresponding section in the first signal.

7.   Apparatus according to claim 6, wherein the synchronising means includes means for identifying the position of one or more delay changes within a section, means for separately synchronising the parts of the section preceding and following each delay change, and means for separately analysing the said parts of the section for distortion.

8.   Method of testing equipment for handling speech signals, comprising the steps of:

supplying a test signal,

receiving a distorted signal, which corresponds to said test signal when, distorted by equipment under test,

selecting individual sections in the test signal and distorted signal, comparing each section in the distorted signal with the corresponding section in the test signal to generate a distortion perception measure which indicates the extent to which the distortion of said section would be perceptible to a human listener, and

combining the results of each such comparison to generate an overall measure of the extent to which the distortion of the signal would be perceptible to a human listener.

9.   Method according to claim 8, wherein the combining step derives an overall measure which takes account of the perceptual importance of each section.

10.   Method according to claim 8 or claim 9 wherein the sections selected are individual utterances, identified by detecting the end of each silent period exceeding a predetermined length.

11.   Method according to claim 10, wherein utterances containing more than a pre-determined length of speech are sub-divided into two or more sections.

12.   Method according to claim 8 or claim 9, wherein individual sections in the signals are selected by detecting whether speech is present in each of a number of arbitrarily chosen sub-divisions of the signal.

13.   Method according to any of claims 8 to 12, wherein the analysis of each section includes the step of synchronising the section in the distorted signal with the corresponding section in the test signal.

5   14.   Method according to claim 13, wherein the section is analysed to identify the position of any delay change, and the parts of the section preceding and following any such delay change are separately synchronised and analysed for degradation.

10 15.   A computer program for performing the steps of any of claims 8, 9, 10, 11, 12, 13 or 14.

16.   A computer program product directly loadable into the internal memory of a computer, comprising software code portions for performing the steps of any of
15 claims 8, 9 , 10, 11, 12, 13 or 14 when said product is run on a computer.

17.   A computer program product stored on a computer usable medium, comprising:
      computer-readable program means for causing a computer to select
20 individual sections in a first signal and in a second signal,
      computer-readable program means for causing the computer to analyse each section in the first signal and the second signal to generate a distortion perception measure which indicates the extent to which the distortion of said section in the second signal as compared with the first signal would be perceptible
25 to a human listener, and
      computer-readable program means for causing the computer to combine the results of each such comparison to generate an overall measure of the extent to which the distortion of the signal would be perceptible to a human listener.
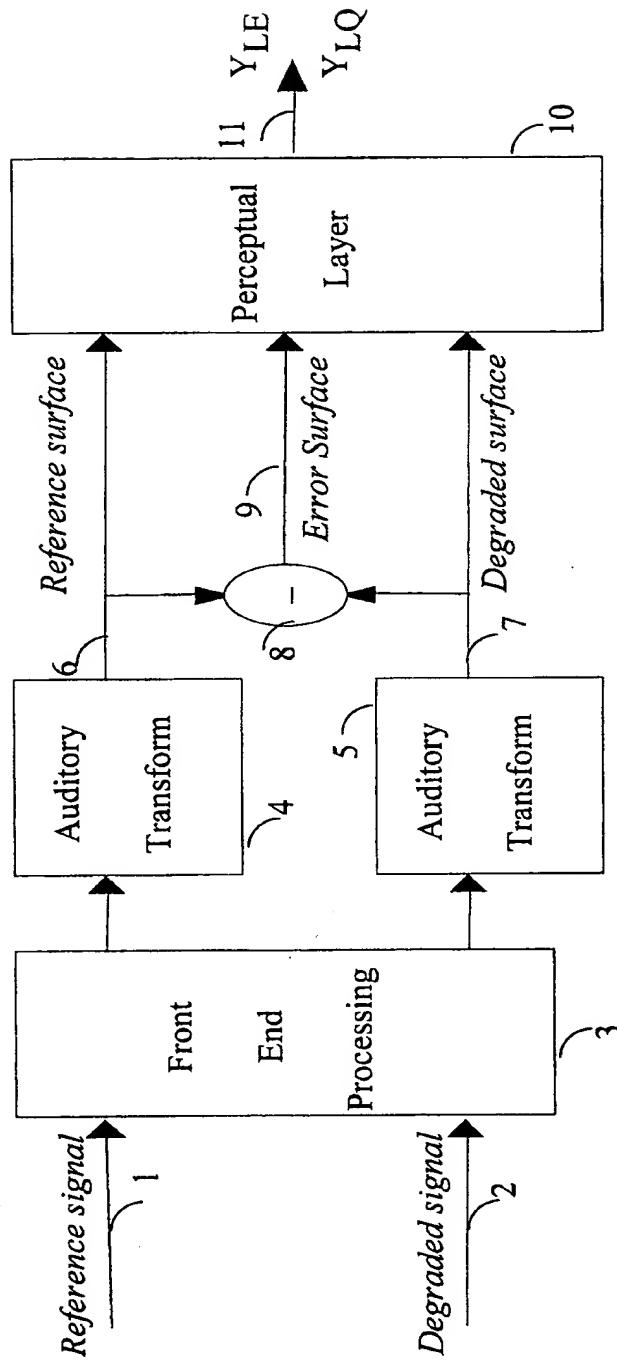
Figure 1

Figure 2

Figure 3

6/7
Auditory transform
of signal [sensation surface]

22 — Threshold

21 — Psychoacoustic loudness map

20 — Power average and downsample

19 — Perceptual Filterbank

*Time & level aligned*
reference or degraded signal
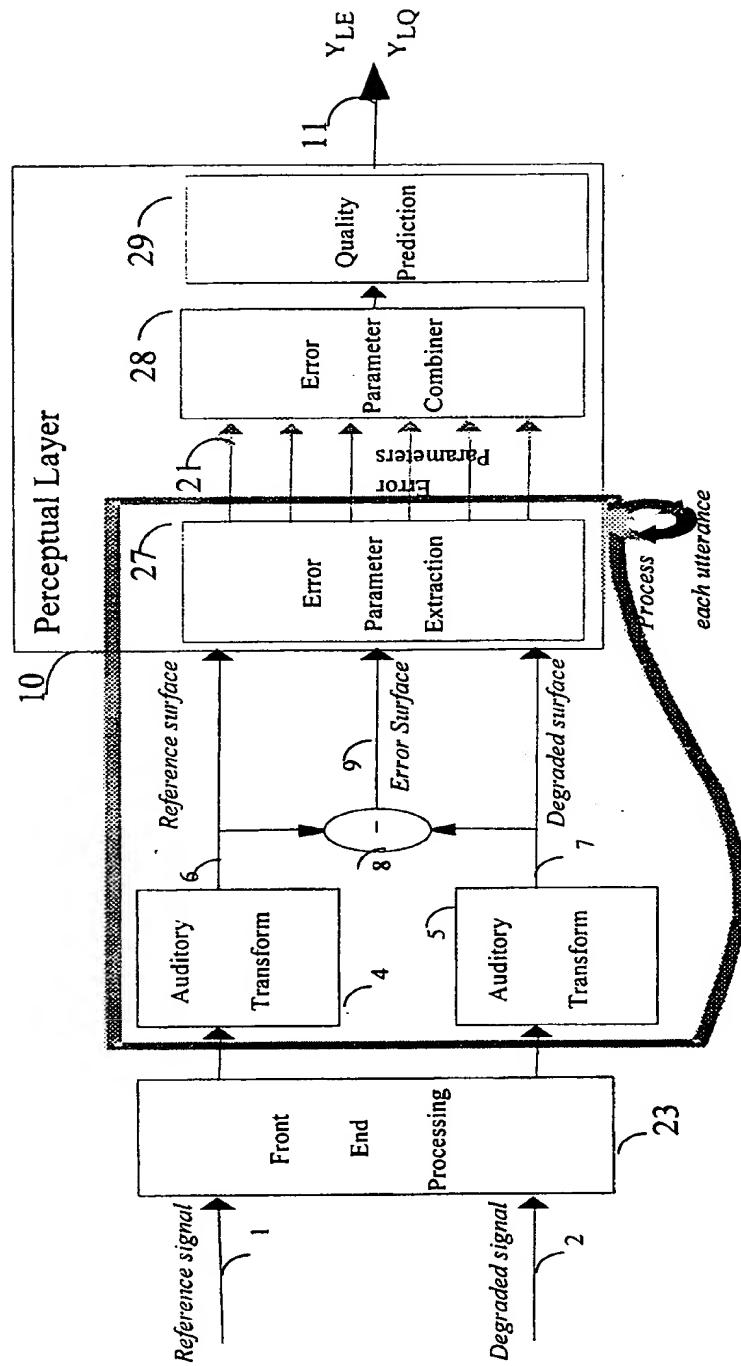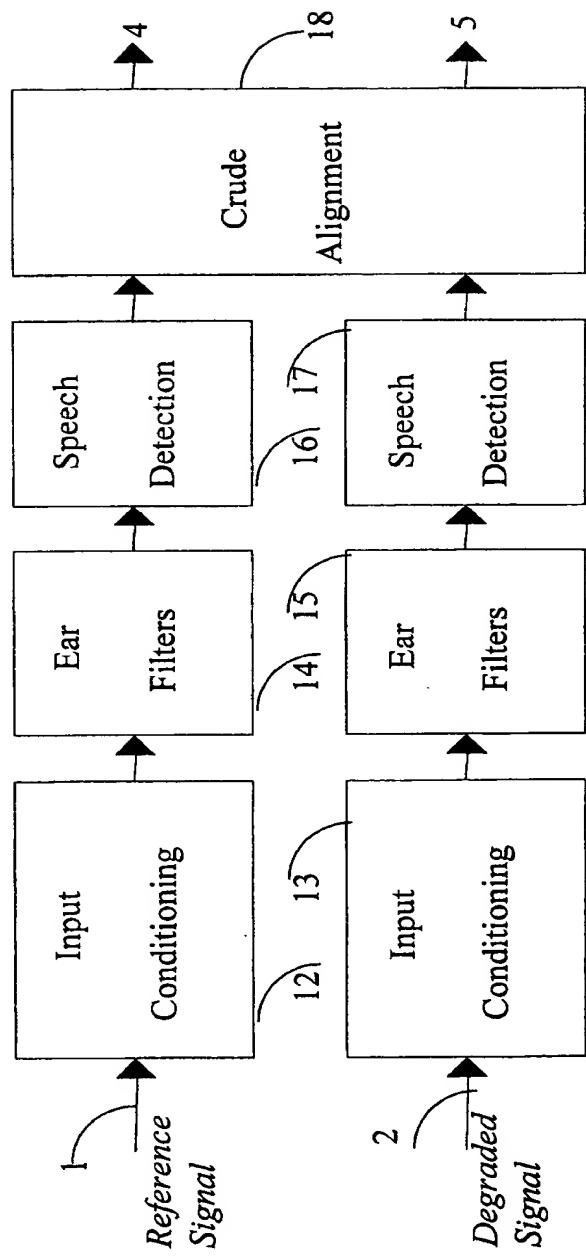
3
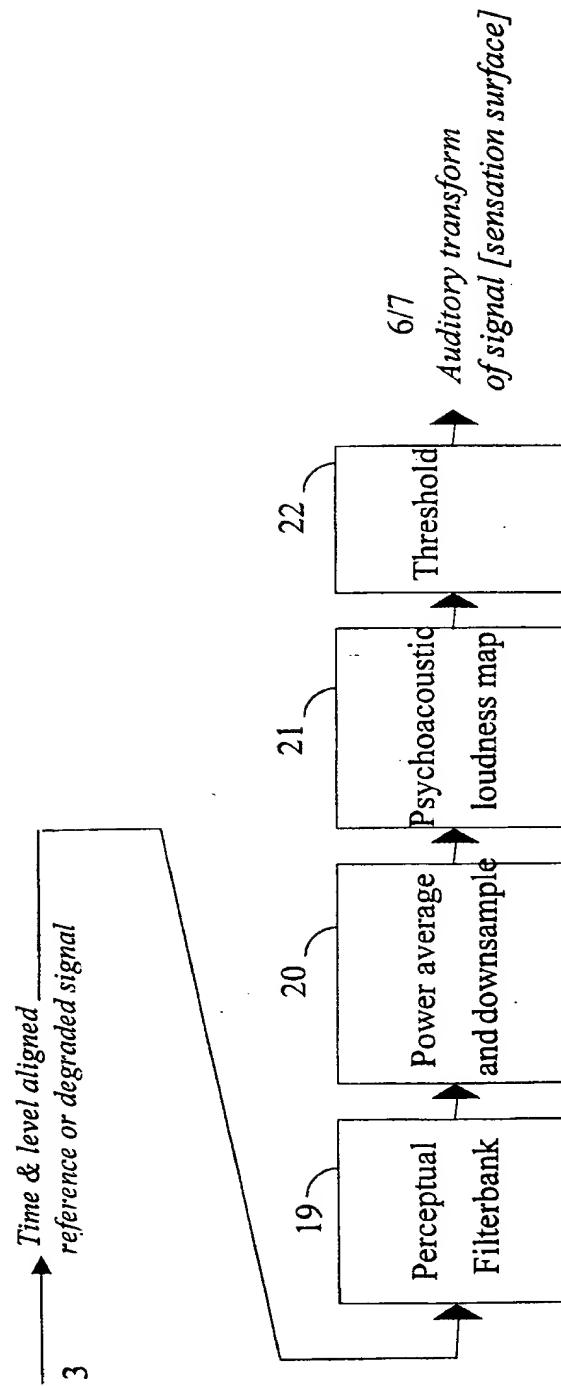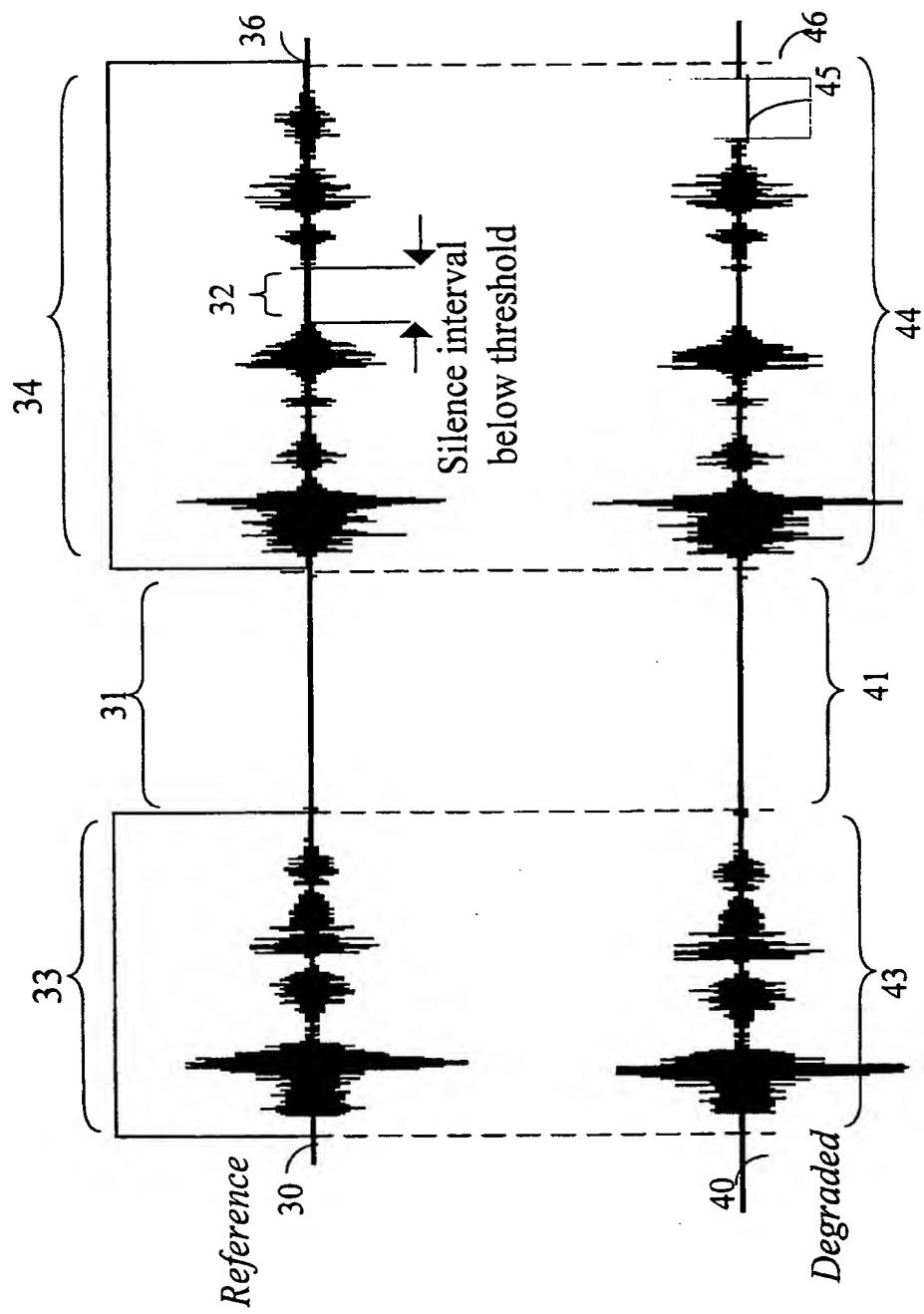
Figure 4

Figure 5

6/8

Figure 6
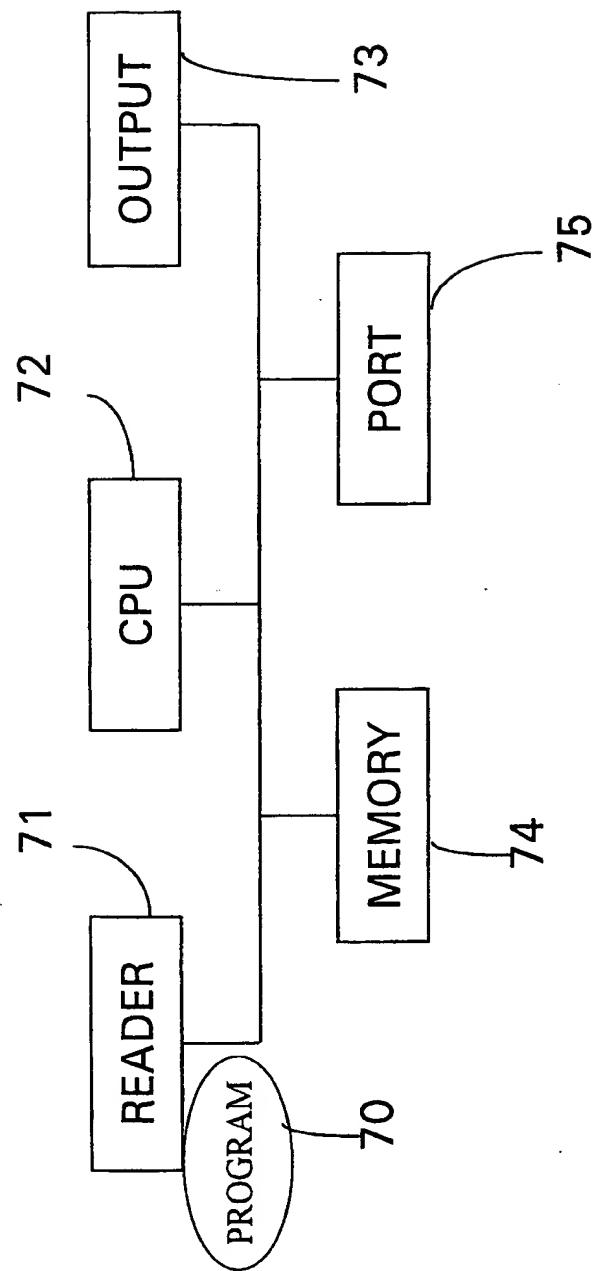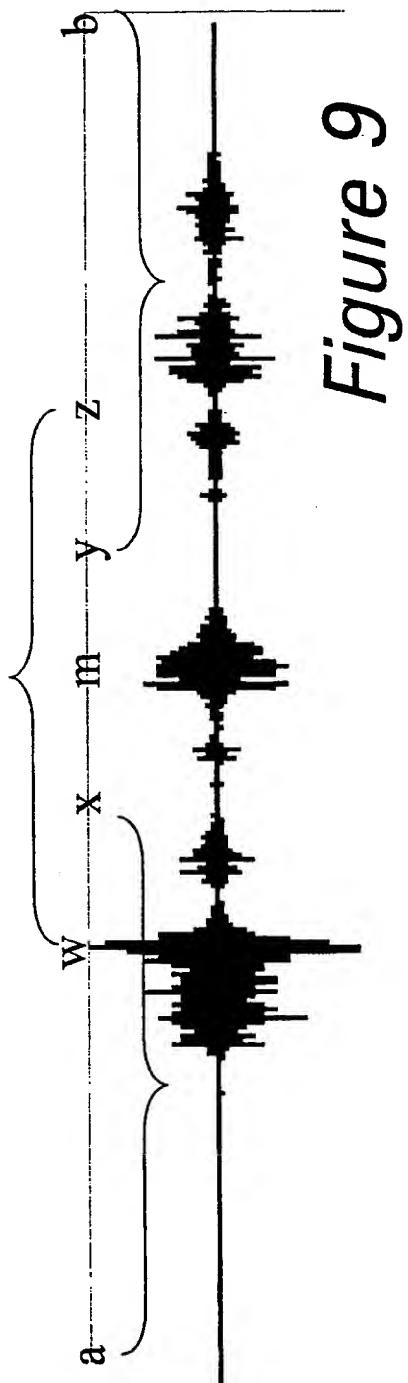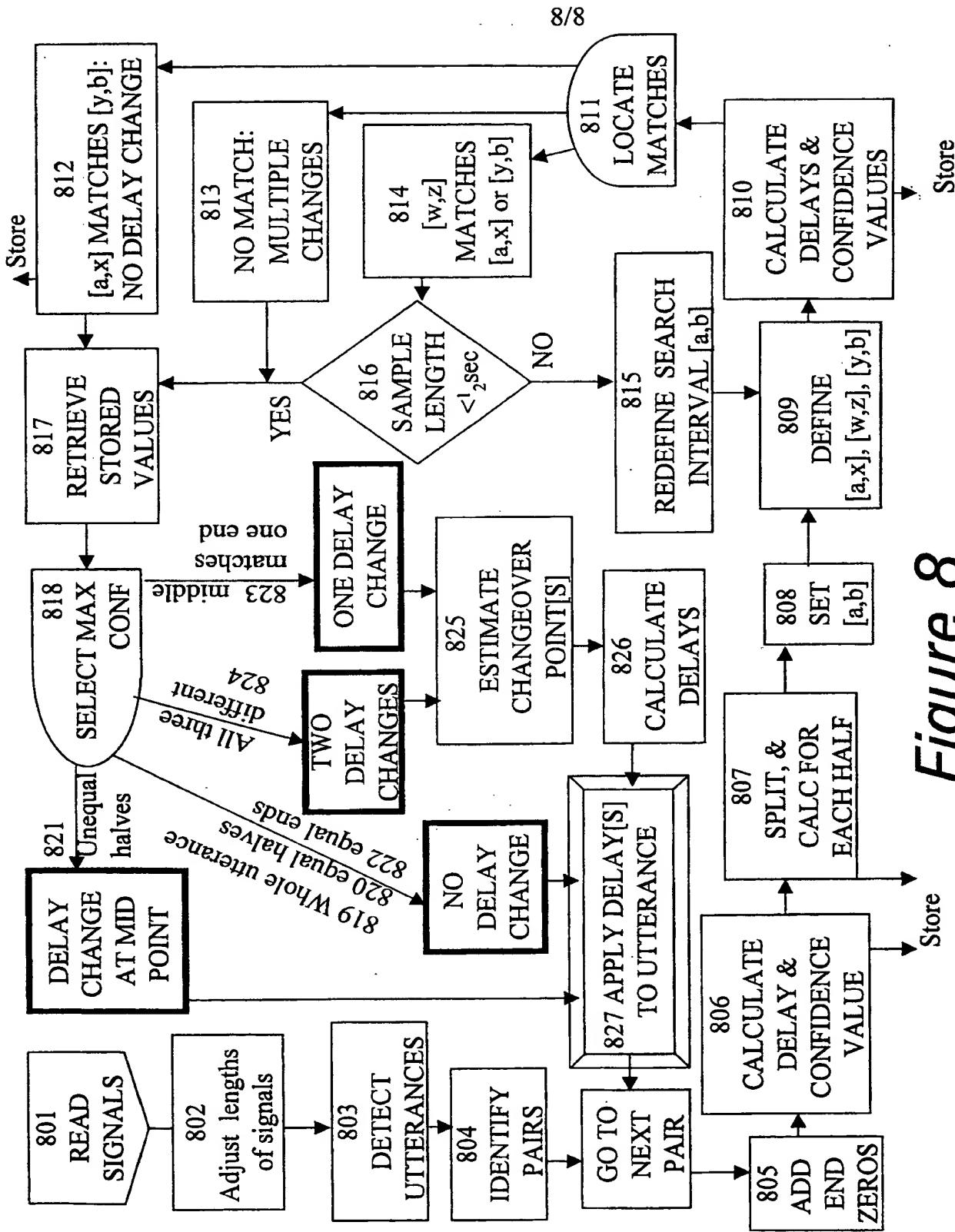
Figure 9

Figure 7

Figure 8

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC 7   H04M3/22   H04B17/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7   H04M   H04B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 647 375 A (BRITISH TELECOMM) 12 April 1995 (1995-04-12) cited in the application page 2, line 1 -page 3, line 6 ---	1,5
A	EP 0 644 674 A (ASCOM INFRASYS AG) 22 March 1995 (1995-03-22) column 4, line 1-40 -----	1,5

| | Further documents are listed in the continuation of box C. | [X] Patent family members are listed in annex. |

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 January 2000	20/01/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Cremer, J

Form PCT/ISA/210 (second sheet) (July 1992)

1

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
EP 0647375	A	12-04-1995	AU	670950 B	08-08-1996
			AU	4350093 A	24-01-1994
			DE	69321590 D	19-11-1998
			DE	69321590 T	01-04-1999
			JP	8501910 T	27-02-1996
			US	5621854 A	15-04-1997
			CA	2137005 A	06-01-1994
			EP	0856961 A	05-08-1998
			ES	2122021 T	16-12-1998
			WO	9400922 A	06-01-1994
			SG	48927 A	18-05-1998
			AU	677401 B	24-04-1997
			AU	6974694 A	17-01-1995
			CA	2161257 A	05-01-1995
			DE	69421704 D	23-12-1999
			EP	0705501 A	10-04-1996
			WO	9501011 A	05-01-1995
			JP	8511672 T	03-12-1996
			SG	47542 A	17-04-1998
			US	5890104 A	30-03-1999
EP 0644674	A	22-03-1995	CH	686752 A	14-06-1996
			FI	944391 A	23-03-1995
			NO	943500 A	23-03-1995